Efficient Variational Bayesian Approximation Method Based on Subspace optimization

Yuling Zheng, Aurélia Fraysse, Thomas Rodet

Efficient Variational Bayesian Approximation Method Based on Subspace optimization

Yuling Zheng, Aurélia Fraysse, Thomas Rodet

Abstract

Variational Bayesian approximations have been widely used in fully Bayesian inference for approximating an intractable posterior distribution by a separable one. Nevertheless, the classical variational Bayesian approximation (VBA) method suffers from slow convergence to the approximate solution when tackling large-dimensional problems. To address this problem, we propose in this paper an improved VBA method. Actually, variational Bayesian issue can be seen as a convex functional optimization problem. The proposed method is based on the adaptation of subspace optimization methods in Hilbert spaces to the function space involved, in order to solve this optimization problem in an iterative way. The aim is to determine an optimal direction at each iteration in order to get a more efficient method. We highlight the efficiency of our new VBA method and its application to image processing by considering an ill-posed linear inverse problem using a total variation prior. Comparisons with state of the art variational Bayesian methods through a numerical example show the notable improved computation time.

Index Terms

Variational Bayesian approximation, subspace optimization, large dimensional problem, unsupervised approach, total variation

I. Introduction

Efficient reconstruction approaches for large dimensional inverse problems involved in image processing are the main concerns of this paper. In general, such problems are ill-posed, which means that the information provided by data is not sufficient enough to give a good estimation of the unknown

Y. Zheng and A. Fraysse are with the Laboratory of signals and systems (CNRS-Supélec-University of Paris-Sud). L2S-SUPELEC, 3, Rue Joliot-Curie, 91190 Gif-sur-Yvette, France. (email: {zheng,fraysse}@lss.supelec.fr)

T. Rodet is with Laboratory of Systems and Applications of Information and Energy Technologies (École Normale Supérieure de Cachan). 61, Avenue du Président Wilson, 94230 Cachan, France. (email: trodet@satie.ens-cachan.fr)

objects. The resolution of ill-posed inverse problems generally relies on regularizations which consist of introducing additional information, see [1] for details. One most commonly used regularization is the Tikhonov one [2]. Nevertheless, Tikhonov regularization leads to an estimator linear with respect to the data. Therefore, its ability to reconstruct non-linear components, such as location and magnitude of jumps or higher order discontinuities, see [3] and [4], is limited.

To overcome such limitations, nonlinear regularizations have been widely used. However, the drawback of those regularizations is the corresponding non quadratic or even non-convex optimization problems which are generally intricate. To tackle this issue, Geman *et al* [3], [4] proposed *half-quadratic* schemes in order to get nonlinear estimates more easily. By introducing auxiliary variables using duality tools, *half-quadratic* schemes transform the original complicated criterion into a half quadratic one where the original variables appear quadratically and the auxiliary variables are decoupled. This *half-quadratic* criterion can be efficiently optimized using classical optimization algorithms, which lead to the desired nonlinear estimates.

In a statistical framework, the half-quadratic schemes have been shown by Champagnat *et al.* [5] as instances of the EM algorithm with latent variables which provide *maximum a posteriori* estimates. Nevertheless, for either the Tikhonov regularization based methods or half-quadratic schemes, only point estimates could be given. Some useful information such as the variance of the estimator, which evaluates its precision, could not be directly obtained. However, such information is accessible if we obtain the posterior distribution of the unknown parameters, which is involved in the Bayesian framework. Furthermore, another advantage of the Bayesian framework is that it provides a systematic way to determine hyperparameters, e.g. a Bayesian hierarchical approach can estimate the hyperparameters as well as unknown parameters by introducing hyperpriors, [6] [7]. Such approaches are known as unsupervised approaches in the literature. In order to exploit these advantages, in the following we are mainly interested in the development of efficient unsupervised Bayesian reconstruction approaches. Nevertheless, the difficulty met in general is that one could only acquire a posterior distribution whose partition function is unknown due to an intractable integral. In such a case, the main challenge is to retrieve the posterior distribution.

In this context, two main types of approaches are employed, stochastic approaches and analytic approximations. Stochastic approaches are based on Markov Chain Monte Carlo (MCMC) techniques [8] which provide an asymptotically exact numerical approximation of the true posterior distribution. The main drawbacks of such approaches are the high computational cost and the poor performance for large dimensional problems involving complicated covariance matrices. The use of such approaches for

large dimensional problems is therefore limited.

Concerning analytic methods, MacKay in [9], see also [10] for a survey, proposed the variational Bayesian approximation (VBA) which aims to determine analytic approximations of the true posterior distribution. In this case, the objective is to find a simpler probability density function (pdf), generally separable, which is as close as possible to the true posterior distribution in the sense of minimizing the Kullback-Leibler divergence. This problem can be formulated as a convex infinite-dimensional optimization problem, whose resolution results in an optimal analytic approximation. However, this approximation does not have an explicit form except for extremely simple cases. In practice, it is generally approached by cyclic iterative methods which update at each iteration one component of the separable distribution while fixing the other ones. Such optimization procedure is known to be time consuming in general. The classical Bayesian methodology is thus not efficient enough when dealing with very large dimensional problems.

In order to obtain more efficient variational Bayesian approaches, a different method has been recently introduced in [11]. It is based on the adaptation of the exponentiated gradient algorithm [12] into the space of pdf, which is no longer a Hilbert space. Instead of approximating an analytical solution of the involved functional optimization problem, this method seeks an approximate solution of this problem iteratively thanks to a gradient-type algorithm with explicit update equations. The optimization of the components of the separable approximation can thus be performed in parallel, which leads to a significant acceleration compared to the classical VBA method.

In order to further improve the method of [11], a natural idea is to consider a new descent direction. In this context, we propose to adapt the subspace optimization methods [13]–[16], into the space of pdf involved in variational Bayesian methodology. The advantage of subspace optimization is its generalized descent directions where Hilbert structure is not required. Moreover, the descent direction can be freely chosen in a subspace of dimension greater than one. This flexibility allows subspace optimization methods to be generally more efficient than conjugate gradient methods [17]. Based on the subspace optimization methods, an improved variational Bayesian optimization method is proposed in this paper.

Moreover, we also consider the application of our improved variational Bayesian method to ill-posed linear inverse problems in image processing. In this context, the total variation (TV) regularization has been popular [18], [19] due to its ability to describe piecewise smooth images. Nevertheless, it is difficult to integrate the TV based prior into the development of unsupervised Bayesian approaches since its partition function depends on hyperparameters and it is not explicitly known. To tackle this problem, Bioucas-Dias *et al.* [20], [21] proposed a closed form approximation of this partition function. Thanks to

this approximation, Babacan *et al.* [22] developed its TV based unsupervised Bayesian approach using the classical VBA method. In this work, we also take advantage of this approximate TV prior. With this prior, we develop our unsupervised approach using the proposed VBA method.

In order to evaluate the proposed approach, we provide also numerical comparisons with [22] which is based on the classical VBA method, and with another approach employing the gradient-type variational Bayesian algorithm proposed in [11]. These comparisons are based on an implementation on a super-resolution problem [23] which aims to reconstruct a high resolution image from several low resolution ones representing the same scene. Moreover, in our reconstruction, we assume that motions between the low resolution images and a reference one could either be estimated in advance or be known through other sources of information. Such configuration appears for instance in astronomy [24] and medical imaging [25].

The rest of this paper is organized as follows: in Section II, we develop our proposed variational Bayesian optimization algorithm. Next, an application of the proposed algorithm to a linear inverse problem is shown in Section III whereas results of numerical experiments on super-resolution problems are given in Section IV. Finally, a conclusion is drawn in Section V.

II. Exponentiated subspace-based variational Bayesian optimization algorithm

A. Notations

In the following $\mathbf{y} \in \mathbb{R}^M$ and $\mathbf{w} \in \mathbb{R}^J$ denote respectively the data vector and the unknown parameter vector to be estimated whereas $p(\mathbf{w})$, $p(\mathbf{w}|\mathbf{y})$ and $q(\mathbf{w})$ represent the prior distribution, the true posterior law and the approximate posterior distribution, respectively.

B. Statement of the problem

The central idea of variational Bayesian methods is to approximate the true posterior distribution by a separable one

$$q(\mathbf{w}) = \prod_{i=1}^{P} q_i(\mathbf{w}_i), \tag{1}$$

where $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_P)$. Here $(\mathbf{w}_i)_{i=1,\ldots,P}$ denote the P disjoint subsets of elements of $\mathbf{w}$ with P an integer between 1 and J.

The optimal approximation is determined by minimizing a measure of dissimilarity between the true posterior distribution and the approximate one. A natural choice for the dissimilarity measure is the

Kullback-Leibler divergence ($\mathcal{KL}$ divergence), see [26]:

$$\mathcal{KL}[q\|p(\cdot|\mathbf{y})] = \int_{\mathbb{R}^J} q(\mathbf{w}) \log \frac{q(\mathbf{w})}{p(\mathbf{w}|\mathbf{y})} d\mathbf{w}. \tag{2}$$

In fact, direct minimization of $\mathcal{KL}[q\|p(\cdot|\mathbf{y})]$ is usually intractable since it depends on the true posterior $p(\cdot|\mathbf{y})$ whose normalization constant is difficult to calculate. However, as given in [27], the logarithm of the marginal probability of the data, called also evidence, can be written as

$$\log p(\mathbf{y}) = \mathcal{F}(q) + \mathcal{KL}[q\|p(\cdot|\mathbf{y})], \tag{3}$$

where $\mathcal{F}(q)$ is the so called negative free energy defined as

$$\mathcal{F}(q) = \int_{\mathbb{R}^J} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} d\mathbf{w}. \tag{4}$$

As $\log p(\mathbf{y})$ is a constant with respect to $q(\mathbf{w})$, minimizing the $\mathcal{KL}$ divergence is obviously equivalent to maximizing the negative free energy. We can see from (4) that for the computation of the negative free energy, the joint distribution $p(\mathbf{y}, \mathbf{w})$ is involved instead of the true posterior law. And this joint distribution can be readily obtained by the product of likelihood and prior distributions. We use hence the negative free energy as an alternative to the $\mathcal{KL}$ divergence.

Let us define a space Ω which is a space of separable pdfs, $\Omega = \{q : \text{pdf and } q = \prod_{i=1}^{P} q_i\}$. Our variational Bayesian problem is thus formulated as

$$q^{opt} = \arg\max_{q \in \Omega} \mathcal{F}(q) \tag{5}$$

Classical variational Bayesian approximation [10] is based on an analytical solution of (5) which is given by $q = \prod_{i=1}^{P} q_i$ with

$$q_i(\mathbf{w}_i) = K_i \exp\left(\langle \log p(\mathbf{y}, \mathbf{w}) \rangle_{\prod_{j \neq i} q_j(\mathbf{w}_j)} \right), \ \forall i = 1, \ldots, P. \tag{6}$$

Here $\langle \cdot \rangle_q = \mathbb{E}_q[\cdot]$ and K_i denotes the normalization constant. We can see from (6) that q_i depends on the other marginal distributions q_j for $j \neq i$. As a result, we cannot obtain an explicit form for q unless in extremely simple cases. Therefore, iterative methods such as the Gauss-Seidel one have to be used to iteratively approach this solution. As a result, classical variational Bayesian method is not efficient enough to treat high dimensional problems.

In this paper, in order to obtain efficient variational Bayesian approaches, instead of firstly giving an analytical solution then iteratively approaching it, we directly propose iterative methods to solve (5) which is a functional optimization problem over functions $(q_i)_{i=1,\ldots,P}$, in the space Ω. As stated in [11],

there exists a problem equivalent to (5) in a separable probability measure space $\mathcal{A} = \bigotimes_{i=1}^{P} \mathcal{A}_i$, the Cartesian product of the $\mathcal{A}_i$, which is defined as follows:

$$\mathcal{A}_i = \{\mu_i : \text{probability measure}$$
$$\text{and } \mu_i(\mathrm{d}\mathbf{w}_i) = q_i(\mathbf{w}_i)\mathrm{d}\mathbf{w}_i \text{ with } q_i \text{ a pdf}\}.$$

Therefore, the optimization of (5) is equivalent to the resolution of the following optimization problem

$$\mu^{opt} = \arg\max_{\mu \in \mathcal{A}} F(\mu), \tag{7}$$

where the functional F satisfies that $\forall \mu \in \mathcal{A}$ of density q, $F(\mu) = \mathcal{F}(q)$.

In [11], the gradient descent method in Hilbert spaces has been transposed into the space of pdfs and, as a result, an exponentiated gradient based variational Bayesian approximation (EGrad-VBA) method whose convergence is proven was proposed to solve the involved functional optimization problem. For the aim of developing more efficient methods, we transpose here in the same context the subspace optimization method which has been shown to outperform standard optimization methods, such as gradient or conjugate gradient methods, in terms of rate of convergence in finite dimensional Hilbert spaces [17].

C. Subspace optimization method in Hilbert spaces

We give in this section a brief introduction of the subspace optimization method in Hilbert spaces. The subspace optimization method has been proposed by Miele *et al.* [13] with a subspace spanned by the opposite gradient and the previous direction. This method is known as Memory Gradient (MG) and can be regarded as a generalization of the conjugate gradient method. More recently, a lot of other subspace optimization methods based on different subspaces, see [15] and [28] for instance, have been proposed. Generally, subspace optimization methods use the following iteration formula:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k = \mathbf{x}^k + \mathbf{D}^k\mathbf{s}^k, \tag{8}$$

where $\mathbf{x}^k$ and $\mathbf{x}^{k+1}$ respectively stands for the estimates at kth and $(k+1)$th iterations, $\mathbf{D}^k = [\mathbf{d}_1^k, \ldots, \mathbf{d}_I^k]$ gathers I directions which span the subspace and the vector $\mathbf{s}^k = [s_1^k, ..., s_I^k]^T$ encloses the step-sizes along each direction. The subspace optimization method offers more flexibility in the choice of the descent direction $\mathbf{d}^k$ by taking linear combinations of directions in $\mathbf{D}^k$.

An overview of existing subspace optimization methods [17] shows that $\mathbf{D}^k$ usually includes a descent direction (e.g. gradient, Newton, truncated Newton direction) and a short history of previous directions. In this work, we consider only the super memory gradient subspace where $\mathbf{D}^k$ is given as follows:

$$\mathbf{D}^k = [-\mathbf{g}^k, \mathbf{d}^{k-1}, \ldots, \mathbf{d}^{k-I+1}]. \tag{9}$$

Here $-\mathbf{g}^k$ is the opposite gradient and $(\mathbf{d}^{k-j})_{j=1,\ldots,I-1}$ are the directions of previous iterations. Chouzenoux *et al.* [17] have addressed a discussion about the dimension of the subspace through simulation results on several image restoration problems. It is shown that in a Hilbert space, for a super memory gradient subspace (9), taking $I = 2$ i.e. a subspace constructed by the opposite gradient and the previous direction, results in the best performance in terms of computation time. In this case, the super memory gradient subspace is degraded to the memory gradient one.

D. Proposed subspace-based variational Bayesian approximation method

In this section, we define our iterative method based on the transposition of the subspace optimization method for the resolution of (7). We use here $k \in \mathbb{N}$, set initially to zero, as the iteration count and assume that μ^k is a Radon probability measure [11] with a density q^k, i.e. $\mu^k(d\mathbf{w}) = q^k(\mathbf{w})d\mathbf{w}$. As we stand in the space of probability density measures, the next iteration should give also a probability density measure absolutely continuous with respect to μ^k. The Radon-Nikodym theorem [29] ensures that this measure should be written as

$$\mu^{k+1}(\mathrm{d}\mathbf{w}) = h^k(\mathbf{w})\mu^k(\mathrm{d}\mathbf{w}), \tag{10}$$

where $h^k \in L^1(\mu^k)$ is a positive function[1]. Since μ^k is a Radon probability measure with a density q^k, we can equivalently write

$$q^{k+1}(\mathbf{w}) = h^k(\mathbf{w})q^k(\mathbf{w}), \tag{11}$$

as updating equation for the approximate density. Furthermore, as we deal with entropy-type functionals, a natural choice for h^k would be an exponential form (see [30]). Moreover, this choice ensures the positivity of pdfs along iterations.

Considering the exponential form of h^k and the subspace optimization principle, we propose

$$h^k(\mathbf{w}) = K^k(\mathbf{s}^k)\exp\left[\mathbf{D}^k(\mathbf{w})\mathbf{s}^k\right], \tag{12}$$

where $K^k(\mathbf{s}^k)$ represents the normalization constant expressed as

$$K^k(\mathbf{s}^k) = \left[\int_{\mathbb{R}^J} \exp\left[\mathbf{D}^k(\mathbf{w})\mathbf{s}^k\right] q^k(\mathbf{w})\mathrm{d}\mathbf{w}\right]^{-1}, \tag{13}$$

and $\mathbf{D}^k(\mathbf{w}) = [d_1^k(\mathbf{w}), \ldots, d_I^k(\mathbf{w})]$ is the set of I directions spanning the subspace. We should state that as we deal with a functional optimization problem, the directions $(d_l^k(\mathbf{w}))_{l=1,\ldots,I}$, are no longer vectors but functions. Finally, $\mathbf{s}^k = [s_1^k, \ldots, s_I^k]^T$ denotes the multi-dimensional step-size.

[1] $h \in L^1(\mu) \Leftrightarrow \int_{\mathbb{R}^J} |h(\mathbf{w})|\mu(\mathrm{d}\mathbf{w}) < \infty$

Due to the exponential form, (12) can also be written as:

$$h^k(\mathbf{w}) = K^k(\mathbf{s}^k)\left[\phi_1^k(\mathbf{w})\right]^{s_1^k}\ldots\left[\phi_I^k(\mathbf{w})\right]^{s_I^k} \tag{14}$$

where $\phi_l^k(\mathbf{w}) = \exp[d_l^k(\mathbf{w})]$, for $l = 1, \ldots, I$.

1) Set of directions spanning the subspace: As discussed in Section II-C, for the super memory gradient subspaces defined in Hilbert spaces (see (9)), the subspace of dimension two, which is known as memory gradient subspace, leads to the most efficient approaches. Moreover, a subspace of dimension two would result in less computation complexity than higher order subspaces. As a result, we consider in this work the transposition of the memory gradient subspace into the space of pdfs. In this case, $I = 2$. The transposition leads to a $\mathbf{D}^k(\mathbf{w})$ consisting of one term related to the Gateaux differential of $\mathcal{F}(q)$ and the other term corresponding to the previous direction. The Gateaux differential of the negative free energy $\mathcal{F}(q)$ is a linear functional defined on Ω by

$$\forall(q,\tilde{q}) \in \Omega^2 \quad \mathrm{d}\mathcal{F}_q(\tilde{q}) = \int_{\mathbb{R}^J} \mathrm{d}f(q,\mathbf{w})\tilde{q}(\mathbf{w})\mathrm{d}\mathbf{w}. \tag{15}$$

In the case of separable q, we have $\mathrm{d}f(q,\mathbf{w}) = \Sigma_i \mathrm{d}_i f(q,\mathbf{w}_i)$, in which $\forall i = 1,\ldots,P$, $\mathrm{d}_i f(q,\mathbf{w}_i)$ is expressed as:

$$\mathrm{d}_i f(q,\mathbf{w}_i) = \langle \log p(\mathbf{y},\mathbf{w})\rangle_{\prod_{j\neq i} q_j(\mathbf{w}_j)} - \log q_i(\mathbf{w}_i) - 1. \tag{16}$$

Mathematically, the structure of our Memory Gradient (MG) subspace is given by

$$\mathbf{D}_{MG}^k(\mathbf{w}) = [\mathrm{d}f(q^k,\mathbf{w}), d^{k-1}(\mathbf{w})], \tag{17}$$

where $\mathrm{d}f(q^k,\cdot)$ is given by (16) and d^{k-1} stands for the direction of the previous iteration, which is given by

$$d^{k-1}(\mathbf{w}) = \log\left(\frac{\prod_i q_i^k(\mathbf{w}_i)}{\prod_i q_i^{k-1}(\mathbf{w}_i)}\right). \tag{18}$$

Our proposed variational Bayesian approximation method based on this subspace is called exponentiated memory gradient subspace-based variational Bayesian approximation in the rest of this paper and is resumed as follows:

Concerning the step size, let us first define

$$g^k : \mathbb{R}^2 \to \mathbb{R} \;\; g^k(\mathbf{s}) = \mathcal{F}\left(K^k(\mathbf{s})q^k(\mathbf{w})\exp\left[\mathbf{D}_{MG}^k(\mathbf{w})\mathbf{s}\right]\right), \tag{19}$$

then the optimal step-size is given by

$$(\hat{\mathbf{s}}^{opt})^k = \arg\max_{\mathbf{s}\in\mathbb{R}^2} g^k(\mathbf{s}). \tag{20}$$

Algorithm Exponentiated Memory Gradient subspace-based Variational Bayesian Approximation (EMG-VBA)

1) Initialize($q^0 \in \Omega$)
2) repeat
   a. determine subspace $\mathbf{D}_{MG}^k(\mathbf{w})$ using (16), (17) and (18)
   b. determine step-sizes $\mathbf{s}^k$
   c. compute
   $q^{k+1}(\mathbf{w}) = K^k\left(\mathbf{s}^k\right) q^k(\mathbf{w}) \exp\left[\mathbf{D}_{MG}^k(\mathbf{w})\mathbf{s}^k\right]$

   until convergence

For the proposed EMG-VBA algorithm, we would prove the following proposition.

Proposition 1. *Let $q^0 \in \Omega$ and $\forall k \geq 0$, let the sequence $\{q^k\}_{k\geq 0}$ be defined by*

$$\begin{aligned} q^{k+1}(\mathbf{w}) =& K^k\left(\mathbf{s}^k\right) q^k(\mathbf{w}) \exp\left[\mathbf{D}_{MG}^k(\mathbf{w})\mathbf{s}^k\right] \\ =& K^k\left(\mathbf{s}^k\right) q^k(\mathbf{w}) \\ & \times \exp\left[s_1^k \mathrm{d}f(q^k, \mathbf{w}) + s_2^k d^{k-1}(\mathbf{w})\right]. \end{aligned} \tag{21}$$

If $\mathbf{s}^k$ is the optimal step-size defined in (20) then

$$\{\mathcal{F}(q^k)\}_{k\geq 0} \text{ converges to a maximum of } \mathcal{F}(q).$$

In a previous work [11], it has been proven that for $q^k \in \Omega$, if q^{k+1} is constructed by the EGrad-VBA with the following updating formula:

$$q_{\text{grad}}^{k+1}(\mathbf{w}) = K^k\left(s_1^k\right) q^k(\mathbf{w}) \exp\left[s_1^k \mathrm{d}f(q^k, \mathbf{w})\right], \tag{22}$$

with an optimal step-size defined as:

$$s_1^k = \arg\max_{s_1 \in \mathbb{R}} \mathcal{F}\left(K^k(s_1) q^k(\mathbf{w}) \exp\left[s_1 \mathrm{d}f(q^k, \mathbf{w})\right]\right), \tag{23}$$

then the negative free energy $\mathcal{F}(q)$ increases. More precisely,

$$\begin{aligned} &\mathcal{F}(q_{\text{grad}}^{k+1}) \geq \mathcal{F}(q^k) \\ &\qquad \text{for } q^k \in \Omega \text{ and } q_{\text{grad}}^{k+1} \text{ constructed by (22)} \end{aligned} \tag{24}$$

Comparing (21) and (22), we notice that the EGrad-VBA algorithm can be identified as a special case of our proposed EMG-VBA algorithm with the second step-size s_2 set to zero. Due to the use of optimal

step-sizes, we obtain

$$\begin{aligned}\mathcal{F}(q_{\text{grad}}^{k+1}) &= \max_{s_1\in\mathbb{R}}\mathcal{F}\left(K^k\left(s_1\right)q^k(\mathbf{w})\exp\left[s_1\mathrm{d}f(q^k,\mathbf{w})\right]\right)\\ &= \max_{s_1\in\mathbb{R},s_2=0}g^k(\mathbf{s})\end{aligned} \tag{25}$$

However, in our proposed EMG-VBA, we consider

$$\begin{aligned}\mathcal{F}(q^{k+1}) &= \max_{\mathbf{s}\in\mathbb{R}^2}g^k(\mathbf{s})\\ &\geq \max_{s_1\in\mathbb{R},s_2=0}g^k(\mathbf{s}) = \mathcal{F}(q_{\text{grad}}^{k+1})\end{aligned} \tag{26}$$

Combining (24) and (26), we obtain

$$\mathcal{F}(q^{k+1}) \geq \mathcal{F}(q^k), \tag{27}$$

which shows that for a sequence $\{q^k\}_{k\geq 0}$ constructed by the proposed EMG-VBA, the sequence of negative free energy $\{\mathcal{F}(q^k)\}_{k\geq 0}$ is a increasing one. Furthermore, $\mathcal{F}$ is a concave functional. As a result, the proposed EMG-VBA generates a sequence $\{\mathcal{F}(q^k)\}_{k\geq 0}$ which converges to a maximum of $\mathcal{F}(q)$.

E. The approximate step-size

Generally, it is too expensive to get the optimal step-size. Therefore, in practice most iterative approaches take sub-optimal ones considering the trade-off between computational cost and difference from the optimal step-size. In scalar cases, typical line search approaches determine trade-off steps by trying out a sequence of values until the fulfillment of certain sufficient conditions, such as Wolfe, Goldstein [31]. An extension of these conditions to multi-dimensional cases can be easily obtained. However, such approximate methods in scalar cases have already been known to be time-consuming when the computation of the objective criterion is expensive, which is the case here. Moreover, it is difficult to adjust multi-dimensional step-sizes to satisfy the sufficient conditions. As a result, the extension to multi-dimensional cases can greatly increase the computational cost. Furthermore, the rate of convergence of such line search methods depends closely on parameters controlling the boundaries of chosen conditions and on the starting point of the step-size. A bad choice of these parameters will cause a slow convergence. Therefore, in this paper, we do not consider such approximate methods and take sub-optimal steps calculated as in [11]. Firstly, we take the second order Taylor expansion of $g^k(\mathbf{s})$ at origin as local approximation,

$$\tilde{g}^k(\mathbf{s}) = g^k(\mathbf{0}) + \left(\frac{\partial g^k}{\partial \mathbf{s}}\bigg|_{\mathbf{s}=\mathbf{0}}\right)^T\mathbf{s} + \frac{1}{2}\mathbf{s}^T\left(\frac{\partial^2 g^k}{\partial\mathbf{s}\partial\mathbf{s}^T}\bigg|_{\mathbf{s}=\mathbf{0}}\right)\mathbf{s}, \tag{28}$$

where $\frac{\partial g^k}{\partial \mathbf{s}}\big|_{\mathbf{s}=\mathbf{0}}$ denotes the gradient vector whereas $\frac{\partial^2 g^k}{\partial \mathbf{s}\partial \mathbf{s}^T}\big|_{\mathbf{s}=\mathbf{0}}$ is the Hessian matrix of the function g^k at $\mathbf{s}=\mathbf{0}$. When s_1 and s_2 are small, $\tilde{g}^k(\mathbf{s})$ is a close approximation of $g^k(\mathbf{s})$. Secondly, we compute our sub-optimal steps by maximizing $\tilde{g}^k(\mathbf{s})$ which is quadratic. Assuming that the Hessian matrix $\frac{\partial^2 g^k}{\partial \mathbf{s}\partial \mathbf{s}^T}\big|_{\mathbf{s}=\mathbf{0}}$ is invertible, our sub-optimal steps are given by

$$(\hat{\mathbf{s}}^{subopt})^k = -\left(\frac{\partial^2 g^k}{\partial \mathbf{s}\partial \mathbf{s}^T}\bigg|_{\mathbf{s}=\mathbf{0}}\right)^{-1} \frac{\partial g^k}{\partial \mathbf{s}}\bigg|_{\mathbf{s}=\mathbf{0}}. \tag{29}$$

III. Application to a linear inverse problem with a Total Variation prior

We show in this section an application of the proposed EMG-VBA to ill-posed inverse problems in image processing. Variational Bayesian approaches are widely used to tackle inverse problems where complicated posterior distributions are involved. In the following, we firstly present such a problem which adopts a Total Variation (TV) prior. Then we develop an unsupervised Bayesian approach with the proposed EMG-VBA.

A. Direct model

We consider here a classical linear direct model:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \tag{30}$$

where $\mathbf{y} \in \mathbb{R}^M$ and $\mathbf{x} \in \mathbb{R}^N$ denote respectively data and unknown parameters to be estimated arranged in column lexicographic ordering. The linear operator $\mathbf{A} \in \mathbb{R}^{M\times N}$ is assumed to be known and $\mathbf{n}$ is an additive white noise, assumed to be i.i.d. Gaussian, $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \gamma_n^{-1}\mathbf{I})$, with γ_n as a precision parameter, i.e. the inverse of the noise variance. The direct model (30) and the hypothesis of i.i.d. Gaussian noise allow an easy derivation of the likelihood function:

$$p(\mathbf{y}|\mathbf{x}, \gamma_n) \propto \gamma_n^{M/2} \exp\left[-\frac{\gamma_n \|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2}{2}\right]. \tag{31}$$

B. Image model

In this work, we consider an image model, more precisely a prior distribution of the unknown $\mathbf{x}$, satisfying two main properties. Firstly, it is able to describe the piecewise smoothness property of images. In the literature, total variation has been largely used in various imaging problems including denoising [18], blind deconvolution [32], inpainting [33] and super-resolution [22]. Secondly, we should have some knowledge of its partition function in order to develop unsupervised approaches which sidestep the difficulty of tuning hyperparameters. Both demands described above lead us to the work of Babacan *et*

al. [22], where an unsupervised Bayesian approach using the total variation (TV) prior was developed. The TV prior is given by

$$p(\mathbf{x}|\gamma_p) = \frac{1}{Z_{TV}(\gamma_p)} \exp\left[-\gamma_p TV(\mathbf{x})\right], \tag{32}$$

where $Z_{TV}(\gamma_p)$ is the partition function and

$$TV(\mathbf{x}) = \sum_{i=1}^{N} \sqrt{(\mathbf{D}_h\mathbf{x})_i^2 + (\mathbf{D}_v\mathbf{x})_i^2}. \tag{33}$$

Here $\mathbf{D}_h$ and $\mathbf{D}_v$ represent first-order finite difference matrices in horizontal and vertical directions, respectively.

The major difficulty is that there is no closed form expression for the partition function $Z_{TV}(\gamma_p)$. To overcome this difficulty, Bioucas-Dias *et al.* [20], [21] proposed an approximate partition function

$$Z_{TV}(\gamma_p) \simeq C\gamma_p^{-\theta N}, \tag{34}$$

where C is a constant and θ is a parameter which has to be adjusted in practice to get better results. This analytic approximation results in

$$p(\mathbf{x}|\gamma_p) \simeq \widetilde{p}(\mathbf{x}|\gamma_p) = \widetilde{C}\gamma_p^{\theta N} \exp\left[-\gamma_p TV(\mathbf{x})\right]. \tag{35}$$

Babacan *et al.* [22] adopted this approximate TV prior with $\theta = 1/2$. In such a case, $Z_{TV}(\gamma_p)$ is approximated by $C\gamma_p^{-N/2}$ which corresponds to the partition function of a multivariate normal distribution of a N-dimensional random vector. However, a TV prior is not similar to a Gaussian one. Therefore, this approximation is not close enough. As a result, in this paper, we keep the parameter θ and adjust it in practice.

C. Hyperpriors

The hyperparameters γ_n and γ_p play an important role in the performance of algorithms. In practice, choosing correct hyperparameters is far from a trivial task. Therefore, we prefer to automatically determine their values. This is done by introducing hyperpriors for the hyperparameters. In order to obtain numerically implementable approaches, conjugate hyperpriors are employed. For γ_n and γ_p, we use Gamma distributions,

$$p(\gamma_n) = \mathcal{G}(\gamma_n|\tilde{a}_n, \tilde{b}_n) \tag{36}$$

$$p(\gamma_p) = \mathcal{G}(\gamma_p|\tilde{a}_p, \tilde{b}_p) \tag{37}$$

where for $a > 0$ and $b > 0$

$$\mathcal{G}(z|a,b) = \frac{b^a}{\Gamma(a)} z^{a-1} \exp(-bz) \tag{38}$$

As we do not have any prior information about γ_n and γ_p, in practice, we consider $\tilde{a}_n = 0$, $\tilde{b}_n = 0$ and $\tilde{a}_p = 0$, $\tilde{b}_p = 0$, which lead to non-informative Jeffreys' priors.

Consequently, we obtain a joint distribution as follows

$$\begin{aligned} p(\mathbf{y},\mathbf{x},\gamma_n,\gamma_p) \widetilde{\propto} \gamma_n^{M/2} \exp\left[-\frac{\gamma_n \|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2}{2}\right] \\ \times \gamma_p^{\theta N} \exp\left[-\gamma_p \sum_{i=1}^{N} \sqrt{(\mathbf{D}_h\mathbf{x})_i^2 + (\mathbf{D}_v\mathbf{x})_i^2}\right] \gamma_n^{-1}\gamma_p^{-1} \end{aligned} \tag{39}$$

where $\widetilde{\propto}$ means "is approximately proportional to". The posterior distribution $p(\mathbf{x},\gamma_n,\gamma_p|\mathbf{y})$ is not known explicitly since the normalization constant is not calculable. In order to proceed the statistic inference of the unknown variables, we resort to the variational Bayesian approximation methods which aims at getting the best separable analytical approximation. In the context of variational Bayesian methods, in order to get numerically implementable approaches, conjugate priors are needed to ensure that each posterior distribution belongs to a given family. Consequently, the optimization of the approximate distribution can be reduced to a numerical approximation of its parameters. Nevertheless, the TV prior introduced above is not conjugate with the likelihood (see (31)) which is a Gaussian distribution. To tackle this problem, Minorization-Maximization (MM) techniques [34] are employed here to get a conjugate variant, as done by Babacan *et al.* in [22].

D. Application of variational Bayesian approximation methods

Let Θ denote the set of all unknown parameters: $\Theta = \{\mathbf{x},\gamma_n,\gamma_p\}$, the objective of variational Bayesian approximations is to give a tractable approximation q_Θ of the true posterior distribution $p(\cdot|\mathbf{y})$. As the TV prior is not conjugate with the likelihood, it is difficult to carry out the maximization of the negative free energy with respect to q_Θ. This difficulty has been solved by adopting Minorization-Maximization (MM) techniques [34], in which maximizing the negative free energy is substituted by maximizing a tractable lower bound. To get such a lower bound of the negative free energy, a lower bound of the approximate TV prior is firstly constructed by introducing positive auxiliary variables $\boldsymbol{\lambda} = [\lambda_1,\ldots,\lambda_N]$, see [22] for details,

$$\begin{aligned} \widetilde{p}(\mathbf{x}|\gamma_p) \geq & M(\mathbf{x},\gamma_p|\boldsymbol{\lambda}) = c\gamma_p^{\theta N} \\ & \times \exp\left[-\gamma_p \sum_{i=1}^{N} \frac{(\mathbf{D}_h\mathbf{x})_i^2 + (\mathbf{D}_v\mathbf{x})_i^2 + \lambda_i}{2\sqrt{\lambda_i}}\right]. \end{aligned} \tag{40}$$

From (40), we can see that the lower bound of the TV prior, $M(\mathbf{x}, \gamma_p|\boldsymbol{\lambda})$, is proportional to a Gaussian distribution and is therefore conjugate to the likelihood. Combining (4) and (40), a lower bound of the negative free energy can be derived as

$$\begin{aligned}\mathcal{F}(q_\Theta) &\geq \mathcal{F}^L(q_\Theta, \boldsymbol{\lambda}) \\ &= \int q_\Theta(\mathbf{x}, \gamma_n, \gamma_p) \log\left(\frac{L(\mathbf{x}, \gamma_p, \gamma_n, \mathbf{y}|\boldsymbol{\lambda})}{q_\Theta(\mathbf{x}, \gamma_n, \gamma_p)}\right) d\mathbf{x} d\gamma_n d\gamma_p, \end{aligned} \tag{41}$$

where

$$L(\mathbf{x}, \gamma_p, \gamma_n, \mathbf{y}|\boldsymbol{\lambda}) = p(\mathbf{y}|\mathbf{x}, \gamma_n) M(\mathbf{x}, \gamma_p|\boldsymbol{\lambda}) p(\gamma_n) p(\gamma_p) \tag{42}$$

is a lower bound of the joint distribution.

Hence the resolution of the problem (5) is performed by alternating the two following steps: maximizing the lower bound $\mathcal{F}^L$ with respect to the pdf q_Θ and updating the auxiliary variable $\boldsymbol{\lambda}$ in order to maximize $\mathcal{F}^L$. Assuming that

$$\begin{aligned} q_\Theta(\Theta) &= q_{\mathbf{x}}(\mathbf{x}) q_{\gamma_n}(\gamma_n) q_{\gamma_p}(\gamma_p) \\ &= \prod_i q_i(x_i) q_{\gamma_n}(\gamma_n) q_{\gamma_p}(\gamma_p), \end{aligned} \tag{43}$$

we carry out an alternate optimization of $\mathcal{F}^L$ with respect to $q_{\mathbf{x}}$, q_{γ_n}, q_{γ_p} and $\boldsymbol{\lambda}$. Altogether, we perform the following alternate iterative scheme[2]:

$$q_{\mathbf{x}}^{k+1} = \arg\max_{q_{\mathbf{x}}} \mathcal{F}^L\left(q_{\mathbf{x}} q_{\gamma_n}^k q_{\gamma_p}^k, \boldsymbol{\lambda}^k\right), \tag{44}$$

$$\boldsymbol{\lambda}^{k+1} = \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}^N} \mathcal{F}^L\left(q_{\mathbf{x}}^{k+1} q_{\gamma_n}^k q_{\gamma_p}^k, \boldsymbol{\lambda}\right) \tag{45}$$

$$q_{\gamma_n}^{k+1} = \arg\max_{q_{\gamma_n}} \mathcal{F}^L\left(q_{\mathbf{x}}^{k+1} q_{\gamma_n} q_{\gamma_p}^k, \boldsymbol{\lambda}^{k+1}\right), \tag{46}$$

$$q_{\gamma_p}^{k+1} = \arg\max_{q_{\gamma_p}} \mathcal{F}^L\left(q_{\mathbf{x}}^{k+1} q_{\gamma_n}^{k+1} q_{\gamma_p}, \boldsymbol{\lambda}^{k+1}\right) \tag{47}$$

The functional optimizations with respect to $q_{\mathbf{x}}$, q_{γ_n} and q_{γ_p} (given by (44), (46) and (47), respectively) are solved by variational Bayesian approximation methods. Since the conditional posterior $p(\gamma_n, \gamma_p|\mathbf{x}, \mathbf{y})$ is separable, it could be approximated efficiently thanks to the classical VBA. In fact, the proposed EMG-VBA is only adopted to approximate the posterior distribution of $\mathbf{x}$ where it improves the rate of convergence. As regards the optimization of the auxiliary variable $\boldsymbol{\lambda}$ (given by (45)), it involves a classical optimization in a Hilbert space.

[2]The auxiliary variable $\boldsymbol{\lambda}$ is chosen to be updated before q_{γ_n} and q_{γ_p} in order to get a simpler iteration formula for q_{γ_p}, see [22].

Due to the use of MM techniques, we manage to get a prior for $\mathbf{x}$ conjugate with the likelihood. Moreover, conjugate Gamma priors are chosen for hyperparameters γ_n and γ_p. Therefore, the optimal approximations $(q_i)_{i=1,\ldots,N}$ belong to a Gaussian family whereas the optimal approximate posterior distributions of hyperparameters q_{γ_n} and q_{γ_p} belong to a Gamma one.

$$q_{\mathbf{x}}^k(\mathbf{x}) = \prod_i \mathcal{N}(x_i|(\mathbf{m}_k)_i, (\boldsymbol{\sigma}_k^2)_i), \tag{48}$$

$$q_{\gamma_n}^k(\gamma_n) = \mathcal{G}(\gamma_n|a_{\gamma_n}^k, b_{\gamma_n}^k), \tag{49}$$

$$q_{\gamma_p}^k(\gamma_p) = \mathcal{G}(\gamma_p|a_{\gamma_p}^k, b_{\gamma_p}^k), \tag{50}$$

Therefore, the optimization of approximate distributions can be performed by iteratively updating their parameters.

1) Optimization of $q_{\mathbf{x}}$ using the proposed EMG-VBA: According to (11), (12) and (16) – (18), we get a distribution of $\mathbf{x}$ depending on the step-size $\mathbf{s}$:

$$\begin{aligned} q_{\mathbf{x}}^{\mathbf{s}}(\mathbf{x}) &= K^k(\mathbf{s}) q_{\mathbf{x}}^k(\mathbf{x}) \exp(s_1 \mathrm{d}f(q_{\mathbf{x}}^k, \mathbf{x}) + s_2 d^{k-1}(\mathbf{x})) \\ &= K^k(\mathbf{s}) q_{\mathbf{x}}^k(\mathbf{x}) \prod_i \left(\frac{q_i^r(x_i)}{q_i^k(x_i)}\right)^{s_1} \left(\frac{q_i^k(x_i)}{q_i^{k-1}(x_i)}\right)^{s_2}. \end{aligned} \tag{51}$$

As the lower bound of the negative free energy is involved, the auxiliary function q_i^r is of the form

$$\begin{aligned} q_i^r(x_i) &\propto \exp\left[\left\langle \log L(\mathbf{x}, \gamma_p, \gamma_n, \mathbf{y}|\boldsymbol{\lambda}^k)\right\rangle_{(\prod_{j\neq i} q_j^k) q_{\gamma_n}^k q_{\gamma_p}^k}\right] \\ &\propto \exp\Bigg[- \int \Bigg(\frac{\gamma_n}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 \\ &\qquad + \gamma_p \sum_{i=1}^N \frac{(\mathbf{D}_h\mathbf{x})_i^2 + (\mathbf{D}_v\mathbf{x})_i^2 + \lambda_i^k}{2\sqrt{\lambda_i^k}} \Bigg) \\ &\qquad \times \left(\prod_{j\neq i} q_j^k(x_j)\mathrm{d}x_j\right) q_{\gamma_n}^k(\gamma_n) q_{\gamma_p}^k(\gamma_p) \mathrm{d}\gamma_n \mathrm{d}\gamma_p \Bigg] \\ &\propto \exp\Bigg[- \frac{\langle\gamma_n\rangle^k}{2}\Big(x_i^2 \mathrm{diag}\left(\mathbf{A}^T\mathbf{A}\right)_i - 2x_i\left(\mathbf{A}^T\mathbf{y}\right)_i \\ &\qquad + 2x_i\left(\mathbf{A}^T\mathbf{A}\mathbf{m}_k\right)_i - 2x_i \mathrm{diag}\left(\mathbf{A}^T\mathbf{A}\right)_i (\mathbf{m}_k)_i \Big) \\ &\qquad - \frac{\langle\gamma_p\rangle^k}{2}\Big(x_i^2 \mathrm{diag}\left(\mathbf{D}_h^T\boldsymbol{\Lambda}^k\mathbf{D}_h + \mathbf{D}_v^T\boldsymbol{\Lambda}^k\mathbf{D}_v\right)_i \\ &\qquad + 2x_i\left(\mathbf{D}_h^T\boldsymbol{\Lambda}^k\mathbf{D}_h\mathbf{m}_k + \mathbf{D}_v^T\boldsymbol{\Lambda}^k\mathbf{D}_v\mathbf{m}_k\right)_i \\ &\qquad - 2x_i \mathrm{diag}\left(\mathbf{D}_h^T\boldsymbol{\Lambda}^k\mathbf{D}_h + \mathbf{D}_v^T\boldsymbol{\Lambda}^k\mathbf{D}_v\right)_i (\mathbf{m}_k)_i \Big) \Bigg] \end{aligned} \tag{52}$$

where $\langle z \rangle^k = \mathbb{E}_{q_z^k}(z)$, $\mathbf{\Lambda}^k = \text{Diag}\left(\frac{1}{\sqrt{\lambda_i^k}}\right)$ is a diagonal matrix with $\left(\frac{1}{\sqrt{\lambda_i^k}}\right)_{i=1,\ldots,N}$ as its elements. Moreover, diag $(\mathbf{M})$ is a vector whose entries are the diagonal elements of matrix $\mathbf{M}$.

The computation of q_i^r, for $i = 1, \ldots, N$, shows that each of them can be identified as a Gaussian distribution with mean and variance expressed explicitly by the two following expressions:

$$\begin{aligned}(\boldsymbol{\sigma}_r^2)_i =& \big[\langle\gamma_n\rangle^k \text{diag}\left(\mathbf{A}^T\mathbf{A}\right)_i \\ &+ \langle\gamma_p\rangle^k \text{diag}\left(\mathbf{D}_h^T\mathbf{\Lambda}^k\mathbf{D}_h + \mathbf{D}_v^T\mathbf{\Lambda}^k\mathbf{D}_v\right)_i\big]^{-1}, \end{aligned} \tag{53}$$

$$\begin{aligned}(\mathbf{m}_r)_i =& (\boldsymbol{\sigma}_r^2)_i \Big[\langle\gamma_n\rangle^k \left(\mathbf{A}^T\mathbf{y} - \mathbf{A}^T\mathbf{A}\mathbf{m}_k + \text{diag}(\mathbf{A}^T\mathbf{A}) \circ \mathbf{m}_k\right)_i \\ &- \langle\gamma_p\rangle^k \left(\mathbf{D}_h^T\mathbf{\Lambda}^k\mathbf{D}_h\mathbf{m}_k + \mathbf{D}_v^T\mathbf{\Lambda}^k\mathbf{D}_v\mathbf{m}_k\right)_i \\ &+ \langle\gamma_p\rangle^k \text{diag}\left(\mathbf{D}_h^T\mathbf{\Lambda}^k\mathbf{D}_h + \mathbf{D}_v^T\mathbf{\Lambda}^k\mathbf{D}_v\right)_i (\mathbf{m}_k)_i\Big], \end{aligned} \tag{54}$$

where $\circ$ denotes the Hadamard product between two vectors.

Based on the above results for q_i^r, using (51), we can derive the expression of $q_{\mathbf{x}}^{\mathbf{s}}(\mathbf{x}) = \prod_i q_i^{\mathbf{s}}(x_i)$ where each component $q_i^{\mathbf{s}}(x_i)$ is a Gaussian distribution with mean $(\mathbf{m}_{\mathbf{s}})_i$ and variance $(\boldsymbol{\sigma}_{\mathbf{s}}^2)_i$ satisfying:

$$\boldsymbol{\sigma}_{\mathbf{s}}^2 = \left[\frac{1}{\boldsymbol{\sigma}_k^2} + s_1\left(\frac{1}{\boldsymbol{\sigma}_r^2} - \frac{1}{\boldsymbol{\sigma}_k^2}\right) + s_2\left(\frac{1}{\boldsymbol{\sigma}_k^2} - \frac{1}{\boldsymbol{\sigma}_{k-1}^2}\right)\right]^{-1}, \tag{55}$$

$$\mathbf{m}_{\mathbf{s}} = \boldsymbol{\sigma}_{\mathbf{s}}^2\left[\frac{\mathbf{m}_k}{\boldsymbol{\sigma}_k^2} + s_1\left(\frac{\mathbf{m}_r}{\boldsymbol{\sigma}_r^2} - \frac{\mathbf{m}_k}{\boldsymbol{\sigma}_k^2}\right) + s_2\left(\frac{\mathbf{m}_k}{\boldsymbol{\sigma}_k^2} - \frac{\mathbf{m}_{k-1}}{\boldsymbol{\sigma}_{k-1}^2}\right)\right]. \tag{56}$$

In above equations, we omit all the indication of vector component $(\cdot)_i$ for the sake of clarity. From (55), we can see that $\boldsymbol{\sigma}_{\mathbf{s}}^2$ is equal to the inverse of a linear combination of three terms where the first term is the present inverse variance, the second term comes from the gradient and the third term is caused by the memory of the previous direction. From (56), we can see that $\mathbf{m}_{\mathbf{s}}$ has the same structure. As stated earlier, the EGrad-VBA can be identified as the proposed EMG-VBA with s_2 set to zero which leads to the elimination of the third term in (55) and (56). Because of the extra term (the third term) compared to EGrad-VBA, EMG-VBA can obtain a closer approximation than EGrad-VBA in one iteration.

The previous distribution is still a function of the step size. A sub-optimal step-size defined by (29) in Section II-E is then adopted. As a result, $(\boldsymbol{\sigma}^2)^{k+1} = \boldsymbol{\sigma}_{\hat{\mathbf{s}}^{subopt}}^2$ and $\mathbf{m}^{k+1} = \mathbf{m}_{\hat{\mathbf{s}}^{subopt}}$.

2) optimization of $\boldsymbol{\lambda}$ in Hilbert spaces: The elements of auxiliary vector $\boldsymbol{\lambda}$ are calculated by maximizing the upper bound $\mathcal{F}^L$ with respect to $(\lambda_i)_{i=1,\ldots,N}$. Since $\mathcal{F}^L$ is concave and differentiable with

respect to $(\lambda_i)_{i=1,\ldots,N}$, its maximum is achieved at the critical point which is given by

$$\begin{aligned}\lambda_i^{k+1} =&\mathbb{E}_{q_{\mathbf{x}}^{k+1}}\left[(\mathbf{D}_h\mathbf{x})_i^2+(\mathbf{D}_v\mathbf{x})_i^2\right]\\ =&\mathbb{E}_{q_{\mathbf{x}}^{k+1}}\left[\mathbf{x}^T(\mathbf{D}_h)_i^T(\mathbf{D}_h)_i\mathbf{x}+\mathbf{x}^T(\mathbf{D}_v)_i^T(\mathbf{D}_v)_i\mathbf{x}\right]\\ =&(\mathbf{D}_h\mathbf{m}^{k+1})_i^2+(\mathbf{D}_v\mathbf{m}^{k+1})_i^2\\ &+\text{trace}\left[(\mathbf{D}_h)_i^T(\mathbf{D}_h)_i\boldsymbol{\Sigma}^{k+1})\right]\\ &+\text{trace}\left[(\mathbf{D}_v)_i^T(\mathbf{D}_v)_i\boldsymbol{\Sigma}^{k+1})\right],\end{aligned} \tag{57}$$

where $(\mathbf{D}_h)_i$ and $(\mathbf{D}_v)_i$ represent the ith row of $\mathbf{D}_h$ and $\mathbf{D}_v$, respectively. And $\boldsymbol{\Sigma}^{k+1}=\text{Diag}\left((\boldsymbol{\sigma}^2)^{k+1}\right)$ is the covariance matrix which is diagonal under the separability hypothesis.

3) optimization of q_{γ_n} and q_{γ_p} using classical VBA: These two distributions are computed using (6). More details of the calculus can be found in [22]. The means of Gamma distributions are used as the estimates of hyperparameters which are

$$\langle\gamma_n\rangle^{k+1}=\frac{M}{\mathbb{E}_{q_{\mathbf{x}}^{k+1}}\left[\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2\right]}, \tag{58}$$

$$\langle\gamma_p\rangle^{k+1}=\frac{\theta N}{\sum_{i=1}^N\sqrt{\lambda_i^{k+1}}}. \tag{59}$$

Altogether, the proposed algorithm for inverse problem using TV prior is summed up in Algorithm 1.

Algorithm 1 Proposed unsupervised variational Bayesian approach

1) Initialize parameters of $(q_i^0)_{i=1,\ldots,N}$, $q_{\gamma_n}^0$, $q_{\gamma_p}^0$ and $\boldsymbol{\lambda}$
2) Update means and variances of q_i^{k+1} for $i=1,\ldots,N$
   a. Compute parameters of intermediary functions q_i^r using (53), (54)
   b. Determine the suboptimal step-sizes $(s_1^{subopt},s_2^{subopt})$ using (29)
   c. Update means and variances of q_i^{k+1} using (55), (56)
3) Update auxiliary vector $\boldsymbol{\lambda}^{k+1}$ using (57)
4) Determine the parameters of $q_{\gamma_n}^{k+1}$ and compute its mean by (58)
5) Determine the parameters of $q_{\gamma_p}^{k+1}$ and compute its mean by (59)
6) Go back to 2) until convergence

IV. Experimental evaluation

The performance of the proposed approach (Algorithm 1) is evaluated through an application on a super-resolution (SR) problem. A SR problem is covered by the linear direct model (30) with a system matrix $\mathbf{A}$ gathering the warping, blurring and down-sampling operators. In fact, the main concern of this section is the evaluation of the time efficiency of the new variational Bayesian algorithm EMG-VBA by comparisons with the existing variational Bayesian approximation techniques, classical VBA and EGrad-VBA [11] for the estimation of images. Therefore, we treat here a non-classical super-resolution problem where, for the sake of simplicity, the system matrix $\mathbf{A}$ is assumed to be known, i.e. no motion parameters are estimated. Meanwhile, this assumption would reduce the implementation limitation of a state of the art approach based on classical VBA [22] to large dimensional problems. In the following, at first, we present briefly this state of the art approach based on the classical VBA and another one using the EGrad-VBA, then we show the comparisons of the proposed approach with these two approaches.

A. State of the art approaches

1) A SR approach based on classical VBA [22]: The linear inverse problem with TV image prior treated in Section III has also been treated recently by Babacan *et al.* in [22] in the context of classical super-resolution. In this paper, we suppose that the motion parameters are known in order to make the SR approach of Babacan *et al.* applicable for large-dimensional problems.

Two major differences exist between the work in [22] and our work presented in Section III-D. The first one is that Babacan *et al.* used the classical VBA for the optimization of $q_{\mathbf{x}}$ whereas we adopt the proposed EMG-VBA. The second difference is that we assume that $q_{\mathbf{x}}$ is fully separable whereas Babacan *et al.* supposed that it is not. Due to the non-separability assumption, classical VBA based on (6) yields a multivariate Gaussian distribution for $q_{\mathbf{x}}^k$. Updating the distribution is also equivalent to updating its parameters, i.e. mean $\mathbf{m}^k$ and covariance matrix Σ^k. These two parameters are given by

$$\mathbf{m}^{k+1} = \boldsymbol{\Sigma}^{k+1}\left[\langle\gamma_n\rangle^k \mathbf{A}^T\mathbf{y}\right], \tag{60}$$

$$(\boldsymbol{\Sigma}^{k+1})^{-1} = \langle\gamma_n\rangle^k \mathbf{A}^T\mathbf{A} + \langle\gamma_p\rangle^k\left(\mathbf{D}_h^T\boldsymbol{\Lambda}^k\mathbf{D}_h + \mathbf{D}_v^T\boldsymbol{\Lambda}^k\mathbf{D}_v\right), \tag{61}$$

We can see from (60) that $\mathbf{m}^{k+1}$ depends on the covariance matrix Σ^{k+1} but the computation of Σ^{k+1} needs the inversion of the matrix given by (61). To bypass the matrix inversion, Babacan *et al.* have adopted the conjugate gradient method to iteratively approximate $\mathbf{m}^{k+1}$, which can be inefficient in large

dimensional cases. For the optimization of the auxiliary variable $\boldsymbol{\lambda}$ and the hyperparameters γ_n, γ_p, the same updating equations as (57), (58) and (59) are obtained. In (57) where the covariance matrix $\boldsymbol{\Sigma}^{k+1}$ is needed, it is approximated by a diagonal matrix whose diagonal entries are equal to the inverse of the diagonal elements of $(\boldsymbol{\Sigma}^{k+1})^{-1}$. Generally, this approximation is not a precise one.

2) EGrad-VBA [11] based approach: We present in this section another approach for comparison which treats the same inverse problem. The only major difference between this approach and the proposed one is that the EGrad-VBA using approximate optimal step-sizes defined in a same way as (29) is adopted for the optimization of $q_{\mathbf{x}}$. In fact, EMG-VBA differs from the EGrad-VBA only in the selection of the direction: EMG-VBA takes the memory of previous direction into consideration whereas EGrad-VBA does not. Here, there is no need to show the details of the application of EGrad-VBA into our inverse problem since it is a degenerated version of the EMG-VBA with a subspace $D^k(\mathbf{w})$ consisting of only $\mathrm{d}f(q^k, \cdot)$. Consequently, the updating equations of the mean and variance of $q_{\mathbf{x}}$ could be easily obtained by considering $s_2 = 0$ in (55) and (56). The updating equations for the auxiliary variable $\boldsymbol{\lambda}$ and the hyperparameters γ_n, γ_p are still the same as (57), (58) and (59).

B. Simulation results

In the following, the approach of Babacan *et al.* [22] is named VBA-SR whereas the EGrad-VBA based approach is referred to as EGrad-SR. The objective of SR is to construct a High-Resolution (HR) image from several degraded Low-Resolution (LR) images representing the same scene, i.e. data. In our experiments, four groups of LR images are generated from four standard images, *Testpat* and *Cameraman* of dimension 256×256 and *Lena, Jetplane* of dimension 512×512. During the generation of LR images, a 3×3 uniform blur kernel and a decimation factor of 4 in both horizontal and vertical directions are used. Moreover, we add i.i.d. white Gaussian noises at SNR levels of 5 dB, 15 dB, 25 dB, 35 dB and 45 dB. However, for the sake of simplicity, we treat a problem where LR images are without rotation and motions are supposed to be known. During the reconstruction, we take twelve LR images as data and assume that the convolution kernel, decimation factor and shifting of LR images are all known. Since the decimation factor of 4 is used, the size of a LR image is $\frac{1}{16}$ of that of the HR image. As a result, the size of twelve LR images is smaller than that of the objective HR image.

In fact, all the considered approaches are based on a same Bayesian model and tackle the same optimization problem. Therefore, in general, these approaches lead to similar reconstruction results. The main concern of this comparison is their rate of convergence. To have a fair comparison, we take the same initializations for all the approaches: $\mathbf{m}_0 = \mathbf{A}^T\mathbf{y}$ as the mean and 100 as the variance of HR

image pixels, the initializations of the auxiliary variables and the hyperparameters are computed from $\mathbf{m}_0$ using (57), (58) and (59). Moreover, the parameter θ involved in the partition function of the image prior needs to be adjusted. A set of experiments carried out with different images show that the best results are achieved with $\theta = 1.1$. As a result, we set $\theta = 1.1$ for all our experiments.

As convergence criterion of the VBA-SR we use $\|\mathbf{m}^k - \mathbf{m}^{k-1}\|/\|\mathbf{m}^{k-1}\| < 10^{-5}$, where $\mathbf{m}^k$ and $\mathbf{m}^{k-1}$ represents the estimate of the HR image at kth and $(k-1)$th iteration, respectively. And for EGrad-SR and our approach, they stop when they achieve a HR image of a PSNR value very close (difference within 1‰) to that obtained by VBA-SR.

TABLE I: PERFORMANCE COMPARISON OF VBA-SR [22], EGRAD-SR [11] AND OUR PROPOSED APPROACH IN TERMS OF NUMBER OF ITERATIONS/CPU TIME (IN SECONDS).

Data		PSNR	VBA-SR	EGrad-SR	Proposed
Testpat	5dB	16.71	88/14.9	**108/1.7**	105/2.0
	15dB	20.98	25/4.1	92/1.5	**73/1.4**
	25dB	25.33	24/6.1	185/2.9	**68/1.3**
	35dB	29.98	33/11.5	282/4.2	**101/2.1**
	45dB	32.62	48/20.1	364/5.3	**128/2.6**
Camera -man	5dB	23.23	181/56.9	336/5.0	**254/4.9**
	15dB	28.77	45/7.4	**80/1.3**	80/1.4
	25dB	33.58	23/4.5	106/1.6	**70/1.4**
	35dB	37.15	26/9.1	231/3.4	**75/1.6**
	45dB	40.52	34/14.6	412/6.3	**107/2.0**
Lena	5dB	27.14	108/100.5	312/16.2	**204/15.7**
	15dB	31.23	21/12.8	**98/5.3**	83/6.1
	25dB	34.61	22/16.8	145/7.6	**71/5.4**
	35dB	37.02	26/25.9	264/13.7	**88/6.7**
	45dB	38.30	32/43.6	392/20.5	**106/7.8**
Jetplane	5dB	32.87	123/76.2	**182/9.7**	172/13.0
	15dB	37.36	31/16.2	**90/4.7**	74/5.6
	25dB	41.04	20/16.0	146/7.7	**72/5.4**
	35dB	44.72	25/25.7	366/19.0	**85/6.3**
	45dB	46.76	31/40.0	342/18.2	**104/7.9**

We show in Table I the number of iterations as well as the computation time taken by VBA-SR, EGrad-SR and the proposed approach to obtain HR images of similar PSNR values (fluctuation < 1‰) which are given in the third column of the Table I. All experiments are run using Matlab R2013a on Intel(R)

Core(TM) i7-3770 CPU (3.40 GHz) with 8.0 GB RAM. In Table I, we use bold numbers to indicate the best results, i.e. the shortest computation time for each data. Comparing the computation time, we can see that the proposed approach is much more efficient than the VBA-SR for all the datasets and more efficient than EGrad-SR in most cases, especially in lower noise ones such as cases where SNR $= 25$, 35 and 45 dB. For instance, in the case where SNR $= 5$ dB for the *Cameraman* image, the proposed approach takes 4.9 seconds which is only 9% of 56.9 seconds used by VBA-SR and slightly smaller than 5.0 seconds taken by EGrad-SR, in the case where SNR $= 25$ dB for the *Lena* image, the proposed approach took 5.4 seconds which is nearly 32% of the time needed by VBA-SR (16.8 seconds) and 71% of the time taken by EGrad-SR (7.6 seconds), in the case where SNR $= 45$ dB for the *Jetplane* image, the proposed approach took 7.9 seconds which is only 20% of the time taken by VBA-SR (40 seconds) and 43% of the time used by EGrad-SR (18.2 seconds). On average, the proposed approach is approximately 4 times faster than the VBA-SR and approximately 1.7 times as fast as EGrad-SR. Comparing the number of iterations, the proposed approach generally takes less iterations than the EGrad-SR (only one exception in the case where SNR $= 15$ dB of *Cameraman* image). This result suggests that the introduction of the memory gradient subspace gives better directions and the proposed approach thus needs less iterations than the approach based on the gradient direction. Even though each iteration of the proposed approach takes more computation time than EGrad-SR due to its complexity, the proposed approach is still more efficient than EGrad-SR in most cases thanks to the decrease in the number of iterations. Concerning the VBA-SR, it takes less iterations than the other two approaches. However, each of its iteration takes much more time since it contains an inner loop due to the use of the conjugate gradient method to avoid the direct matrix inversion. As a result, VBA-SR is less efficient than the proposed approach in terms of computation time.

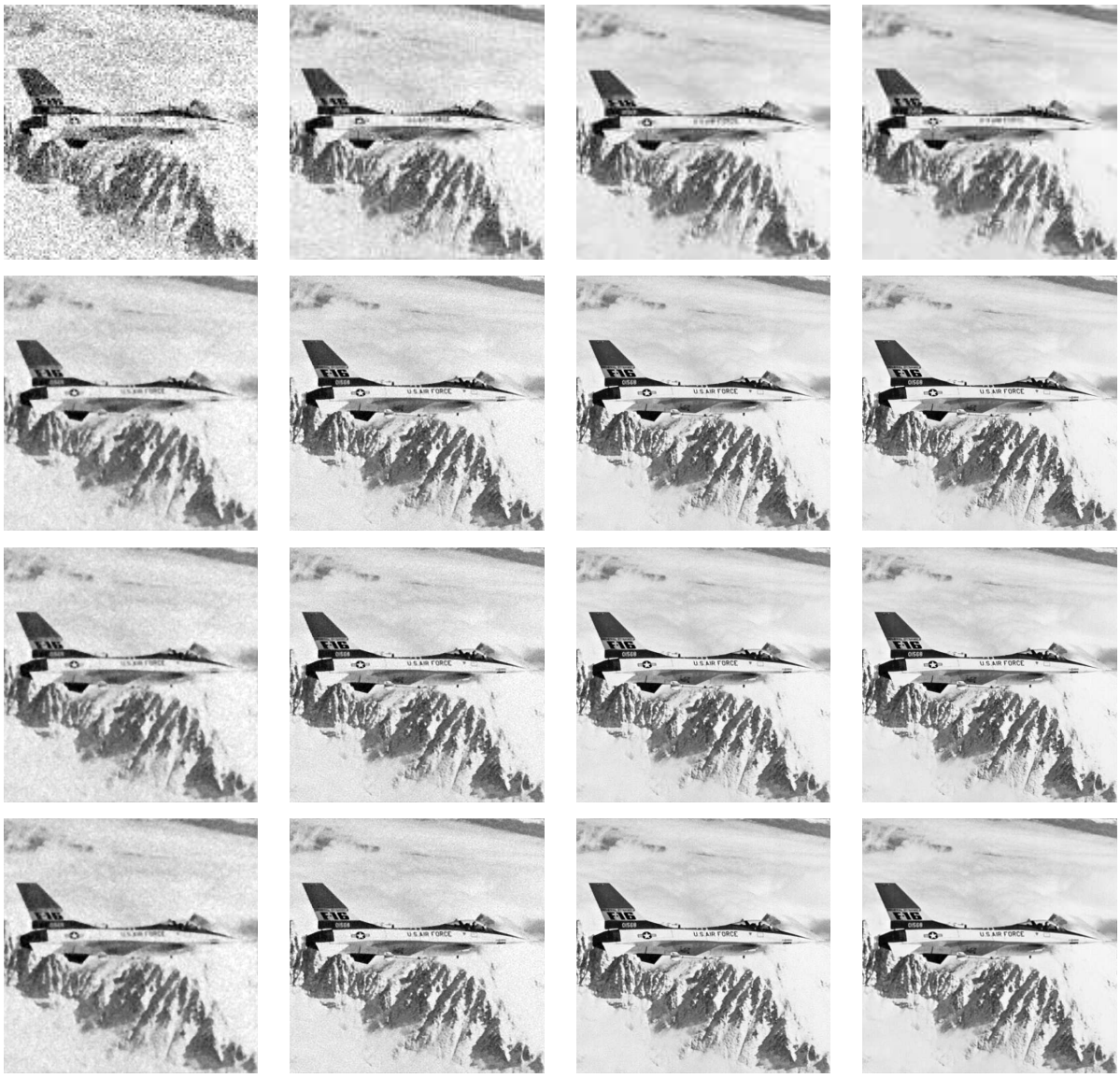
In order to compare the visual quality of reconstructed images, we show in Fig. 1 and Fig. 2 one of the LR images, the reconstructed images obtained by VBA-SR, EGrad-SR and our proposed approach for *Cameraman* and *Jetplane* LR images of SNR $= 5$, 15, 25, 35 dB. Comparing the LR images (shown in the top row) and the reconstructed HR images (given in the second, third and the bottom row) in Fig. 1 and Fig. 2, we can see that all the approaches increase the image resolution. Even in highly noisy cases where SNR $= 5$ dB, the noise present in LR images is effectively reduced meanwhile the image edges are not over-smoothed. Moreover, we can see that all the approaches give very similar HR images in each case, which is coherent with the similar PSNR values achieved by these approaches. Furthermore, as stated above, all the HR images were obtained with a same value of θ. We can see here that our approaches work well with this value in all the tested cases.

(a) (b) (c) (d)

Fig. 1: One of the LR images (top row), HR images obtained by VBA-SR (the second row), EGrad-SR (the third row), Proposed approach (bottom row) for the *Cameraman* in the cases where SNR $=$ (a) 5 dB; (b) 15 dB; (c) 25 dB (d) 35 dB. All images are presented in the same range of grayscale.

V. Conclusion

In this paper, we proposed an efficient variational Bayesian approximation method based on the transposition of the memory gradient subspace algorithm into the space of probability density functions.



(a) (b) (c) (d)

Fig. 2: One of the LR images (top row), HR images obtained by VBA-SR (the second row), EGrad-SR (the third row), Proposed approach (bottom row) for the *Jetplane* in the cases when SNR $=$ (a) 5 dB; (b) 15 dB; (c) 25 dB (d) 35 dB. All images are presented in the same range of grayscale.

This approach is applied to a linear inverse problem where a TV image prior with an approximate partition function and Jeffrey's hyperpriors are used, which results in a fully automatic algorithm. We have shown on a super-resolution problem that the proposed algorithm is much more fast than state of art approaches.

The reason is that we have integrated the memory gradient subspace optimization method which allows more flexibility in the choice of directions.

REFERENCES

[1] G. Demoment, "Image reconstruction and restoration: Overview of common estimation structure and problems," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, no. 12, pp. 2024–2036, Dec. 1989.

[2] A. Tikhonov, "Regularization of incorrectly posed problems," *Soviet. Math. Dokl.*, vol. 4, pp. 1624–1627, 1963.

[3] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367–383, Mar. 1992.

[4] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 932–946, Jul. 1995.

[5] F. Champagnat and J. Idier, "A connection between half-quadratic criteria and EM algorithms," *IEEE Signal Process. Lett.*, vol. 11, no. 9, pp. 709–712, Sep. 2004.

[6] R. Molina, "On the hierarchical Bayesian approach to image restoration: Application to astronomical images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 11, pp. 1122–1128, nov 1994.

[7] D. M. Higdon, J. E. Bowsher, V. E. Johnson, T. G. Turkington, D. R. Gilland, and R. J. Jaszczak, "Fully Bayesian estimation of Gibbs hyperparameters for emission computed tomography data," *IEEE Trans. Med. Imag.*, vol. 16, no. 5, pp. 516–526, oct 1997.

[8] C. P. Robert and G. Casella, *Monte-Carlo Statistical Methods*, ser. Springer Texts in Statistics. New York: Springer, 2000.

[9] D. J. C. MacKay, "Ensemble learning and evidence maximization," http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.4083, 1995.

[10] V. Šmídl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Springer, 2006.

[11] A. Fraysse and T. Rodet, "A measure-theoretic variational Bayesian algorithm for large dimensional problems." Tech. Rep., 2012, http://hal.archives-ouvertes.fr/docs/00/98/02/24/PDF/var_bayHAL.pdf.

[12] J. Kivinen and M. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *Information and Computation*, vol. 132, no. 1, pp. 1–63, 1997.

[13] A. Miele and J. Cantrell, "Study on a memory gradient method for the minimization of functions," *Journal of Optimization Theory and Applications*, vol. 3, no. 6, pp. 459–470, 1969.

[14] M. Wolfe and C. Viazminsky, "Supermemory descent methods for unconstrained minimization," *Journal of Optimization Theory and Applications*, vol. 18, no. 4, pp. 455–468, 1976.

[15] G. Narkiss and M. Zibulevsky, *Sequential Subspace Optimization Method for Large-Scale Unconstrained Problems*. Technion-IIT, Department of Electrical Engineering, Oct. 2005.

[16] E. Chouzenoux, A. Jezierska, J. C. Pesquet, and H. Talbot, "A majorize-minimize subspace approach for l2-l0 image regularization," *SIAM Journal on Imaging Sciences*, vol. 6, no. 1, pp. 563–591, 2013.

[17] E. Chouzenoux, J. Idier, and S. Moussaoui, "A Majorize-Minimize strategy for subspace optimization applied to image restoration," *IEEE Trans. Image Process.*, vol. 20, no. 18, pp. 1517–1528, 2011.

[18] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[19] I. Pollak, A. S. Willsky, and Y. Huang, "Nonlinear evolution equations as fast and exact solvers of estimation problems," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 484–498, 2005.

[20] J. Bioucas-Dias, M. Figueiredo, and J. Oliveira, "Adaptive total-variation image deconvolution: A majorization-minimization approach," in *Proc. EUSIPCO*, 2006, pp. 1–4.

[21] J. P. Oliveira, J. M. Bioucas-Dias, and M. A. Figueiredo, "Adaptive total variation image deblurring: a majorization–minimization approach," *Signal Processing*, vol. 89, no. 9, pp. 1683–1693, 2009.

[22] S. Babacan, R. Molina, and A. Katsaggelos, "Variational Bayesian super resolution," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 984–999, 2011.

[23] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *Signal Processing Magazine, IEEE*, pp. 21–36, May 2003.

[24] T. Rodet, F. Orieux, J.-F. Giovannelli, and A. Abergel, "Data inversion for over-resolved spectral imaging in astronomy," *IEEE J. Sel. Top. Sign. Proces.*, vol. 2, no. 5, pp. 802–811, Oct. 2008.

[25] H. Greenspan, G. Oz, N. Kiryati, and S. Peled, "MRI inter-slice reconstruction using super-resolution," *Magnetic Resonance Imaging*, vol. 20, no. 5, pp. 437–446, 2002.

[26] J. Bernardo, "Expected information as expected utility," *The Annals of Statistics*, vol. 7, no. 3, pp. 686–690, 1979.

[27] R. A. Choudrey, "Variational methods for Bayesian independent component analysis," Ph.D. dissertation, University of Oxford, 2002.

[28] Z. Shi and J. Shen, "A new super-memory gradient method with curve search rule," *Applied mathematics and computation*, vol. 170, no. 1, pp. 1–16, 2005.

[29] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York: McGraw-Hill Book Co., 1987.

[30] P. Tseng and D. P. Bertsekas, "On the convergence of the exponential multiplier method for convex programming," *Mathematical Programming*, vol. 60, no. 1-3, pp. 1–19, 1993.

[31] J. Nocedal and S. Wright, *Numerical Optimization*, ser. Series in Operations Research. New York: Springer Verlag, 2000.

[32] T. F. Chan and C. K. Wong, "Total variation blind deconvolution," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 370–375, 1998.

[33] J. h. Shen and T. F. Chan, "Mathematical models for local nontexture inpaintings," *SIAM Journal on Applied Mathematics*, vol. 62, no. 3, pp. 1019–1043, 2002.

[34] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, 2004.